llm-jp-eval-mm: 日本語視覚言語モデルの自動評価基盤

前田 航希 ♣,‡*，杉浦 一瑳 ◇,‡*，小田 悠介 ‡，栗田 修平 ♠,‡，岡崎 直観 ♣,‡
♣ 東京科学大学，◇ 京都大学，♠ 国立情報学研究所，
‡ 国立情報学研究所 大規模言語モデル研究開発センター
{koki.maeda@nlp.,okazaki@}@comp.isct.ac.jp, sugiura.issa.q29@kyoto-u.jp,
{skurita, odashi}@nii.ac.jp

概要

視覚言語モデル（VLM）の研究は急速に進展しているが，日本語の視覚言語（V&L）タスクにおける評価環境は未だ十分に整備されていない．その結果，日本語評価データセットは散逸し，網羅的な性能評価をすることは困難な状態にあった．本研究では，日本語性能に関する複数のマルチモーダル課題を統一した環境で評価するためのツールキット llm-jp-eval-mm を提案する．本ツールキットは既存の 6 つの日本語マルチモーダル課題を統合し，モデル出力を複数の指標で一貫して評価するベンチマーク基盤である．本稿は llm-jp-eval-mm の構築と継続的な開発のための設計概要を述べ，公開されている 13 種類の日本語・多言語 VLM を評価した結果を報告し，既存研究の知見に照らして議論する．

1 はじめに

言語モデル (LM) の開発において，複数のベンチマークを手軽に横断し，性能を一元的かつ再現可能な形で評価できるツールキットは，モデル改善のサイクルを大幅に短縮することに繋がる．実際，LM の日本語性能を評価する基盤 [1] や視覚言語モデル (VLM) の評価環境 [2, 3] が整いつつあり，多様なモデルを同条件下で比較検証する土台が徐々に整備されている．しかし，こうした試みは VLM の日本語性能を評価する課題セットにおいては未だ進まず，モデル開発に伴って提案された小規模なベンチマーク [4, 5] に依拠している．評価条件や指標が統一されない環境では，VLM 開発者が公表する情報同士の直接的な比較ができず，得られる知見が断片的になる懸念がある．したがって，再現性の高い実験や効果的なモデル改良への道標として，日本語 V&L モデルを異なるデータセット上でも共通の基準で横断評価できる評価基盤の早急な整備が必要である．

本研究では，日本語 VLM に対する包括的な性能評価のために，llm-jp-eval-mm を提案する． llm-jp-eval-mm は複数の日本語 V&L タスク [4, 5, 6, 7, 8, 9] の入出力フォーマットを統一し，一貫したプロトコル下でモデル性能を評価できる基盤である．タスク定義とスコア計算処理を独立したクラス構造で実装することで，新たなタスクやモデルへの拡張・適用が容易になる．また，推論部分をモデル開発者側に委ね，評価基盤そのものの保守性を担保した．

llm-jp-eval-mm は Apache License 2.0 の上で公開されており，PyPI[1] および GitHub[2] からインストールできる．本ツールキットは，日本語 VLM の継続的な改良・評価の基盤となり，効率的なモデル開発や学術的知見の蓄積に寄与することが期待される．

2 関連研究

日本語での VLM 開発は，Stability AI が Japanese InstructBLIP Alpha [10] を開発したのを皮切りに進展を続け，LLaVA [11] の学習データを日本語に翻訳して作成された Heron [4] や，LLaVA-1.5 [12] を基に Japanese Stable VLM [13], LLaVA-CALM2-SigLIP [14], LLM-jp-3 VILA [15] が発表された．また，Llama-3-EvoVLM-JP-v2 [6] はモデルマージを行い VLM に日本語能力を獲得させた．さらに，多言語に対応した VLM [16, 17, 18] も提案されており，それらの日本語能力の検証が必要とされている．

一方，公開された VLM の日本語性能を検証するための評価ベンチマークの整備は追いついておらず，モデル開発者がモデル提案時に限定的な評価データセットを独自に構築して性能を報告している状況である．例えば Heron は Japanese Heron Bench [4] を，Llama-3-EvoVLM-JP-v2 は JA-MultiImage-VQA [5]

* 2 人の著者は本研究に等しく貢献した．

1） https://pypi.org/project/eval-mm
2） https://github.com/llm-jp/llm-jp-eval-mm

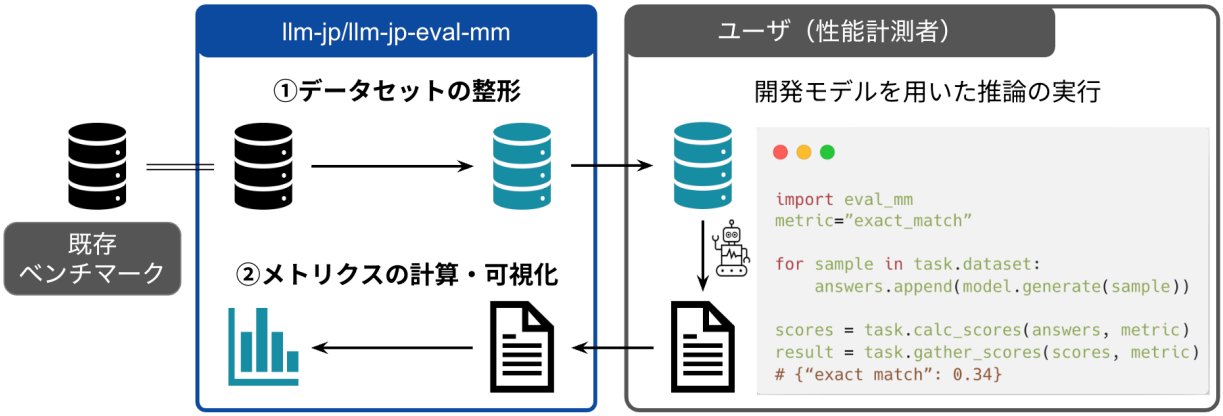

図 1　llm-jp-eval-mm の評価フレームワーク．

を提供して評価スコアを他モデルと比較し，性能を呈示している．そのため，比較対象や評価手法が統一されず，各研究結果を体系的に整理することが困難となっている．直近では JMMMU [9] のように，特定のドメインを扱った比較的大規模な日本語 V&L ベンチマークも提案されている．しかし，単一のベンチマークで総合的な性能を評価することは依然として困難であるため，複数データセットの評価結果を横断的に比較する必要がある．

VLM の評価基盤として，生成タスクの評価では LVLM-eHub [2] のようにモデル同士の出力を直接人間が比較する手法がある．この手法は小規模なデータセット上での緻密な評価には有効だが，人的・時間的コストが高く広範な利用は困難である．lmms-eval [3] は 50 を超える V&L タスクおよび 10 種以上のモデルを網羅する自動評価基盤であり，推論・評価をモジュール化して提供することで，ベンチマークやベースラインモデルの追加を容易にしている．しかしながら，評価設定の大部分を YAML 形式で行い，モデル種別や推論環境に加え，評価設定に合わせて VLM の推論コードを独自に定義している．これが原因で，設定の追加ごとに推論コードの更新を要求するため，評価基盤の継続的な開発にかかるコストが大きい．llm-jp-eval-mm は lmms-eval に追従しつつ，YAML 形式での設定を廃した上で VLM の推論モジュールを切り離して実装することで，簡便さを維持したまま保守性を向上させることを目指した．

3　llm-jp-eval-mm

評価基盤は (1) 継続的な更新 (2) 利用者の学習コスト低減 (3) 評価の正当性の担保 が望まれる．しかし lmms-eval のような実装方法では，モデルごとに異なる推論環境への対応が負担となり，評価基盤の開発側が全モデルの包括的なサポートを継続的に提供することは現実的ではない．特に VLM における推論は画像入力インターフェースや依存ライブラリのバージョン差異による影響が大きく，統一的な推論環境の確立は容易ではない上に，無理な標準化は新たな推論手法やモデル構造の提案を阻害するおそれがある．

こうした課題に対処するため，本研究では日本語の視覚言語タスクにおける性能評価を容易かつ柔軟に行えるツールキット llm-jp-eval-mm を提案する．図 1 に概要を示すように，llm-jp-eval-mm は以下の 2 点に注力することで，日本語視覚言語課題の性能評価を支援する．

- 評価データセットのフォーマット統一
- 推論結果の評価および可視化

本ツールキットでは，評価基盤としてデータセット統合・評価指標計算を担い，推論部分はモデル開発者や利用者に委ねる．この方針により，評価基盤の継続的な更新が簡便化され，モデル利用者が新たなタスク評価を行う際の導入コストも低減される．また，モデル開発者は自身のモデルごとに必要な前処理・後処理や依存するライブラリの変更を行いやすくなり，評価環境への統合が容易になる．参考実装として，いくつかのベースラインモデルに対する統一的条件下での推論コードをツールキットと併せて公開する．

llm-jp-eval-mm は本稿執筆時点で 6 つの画像質問応答課題に対応し（表 1），幅広いドメインでモデルを評価可能なセットを提供する．

llm-jp-eval-mm の設計概要　本ツールキットは，評価タスクを統括する `Task` クラスと，各種評価指標

表 1 llm-jp-eval-mm が提供する評価データセットの一覧.†: 一部の画像のライセンスが不明である.

評価タスク名	サンプル数	画像数	質問あたり画像数	LICENSE	評価指標（デフォルト）
Japanese Heron Bench [4]	103	21	1	CC BY 1.0 / CC BY 2.0†	LLM-as-a-judge
JA-VLM-Bench-In-the-Wild [6]	50	42	1	Unsplash License	ROUGE-L
JA-VG-VQA500 [7]	4,000	500	1	CC BY 4.0	ROUGE-L
JA-Multi-Image-VQA [5]	55	135	2≤	Apache 2.0	ROUGE-L
JDocQA [8]	1,176	551	1≤	CC BY-SA 4.0	Exact Match / BLEU
JMMMU [9]	1,320	1,326	1≤	Apache 2.0	Exact Match

表 2 評価実験におけるモデルの推論設定.

max_new_tokens	256	top_p	1.0
num_beams	1	do_smaple	False
temperature	0.0		

計算を担う Scorer クラスで構成されている．Task クラスは特定のベンチマークに対応し，評価データセットのダウンロードや整形，及びモデル推論結果を用いた評価処理への外部インターフェースを提供する．Scorer クラスは特定の評価指標に対応し，対象文と参照文の比較及び評価スコア・統計値の算出を実装する．各 Task は指定された評価指標に応じた Scorer を呼び出し，評価指標を算出する．これは HELM [19] のように複数評価指標で評価することを見越した設計である．各設問ごとのスコア計算 (score()) と設問全体のスコアの集約計算 (aggregate()) を分離し，単体テストの簡便化を図っている．

実験の再現性 公正なモデル性能評価には，プロンプトやハイパーパラメータなど推論条件の統一が必要である [20]．llm-jp-eval-mm では，VLM 開発者が新たなモデル推論コードをプルリクエストで提供することで，公式ベースラインモデル群に組み込まれる仕組みを用意している．これにより，モデル開発者はコミュニティ標準の評価環境で成果を共有し，再現性確保へのインセンティブを得られる．これは拡張性や公正性を高めると同時に，開発者コミュニティによる積極的な再現性改善への関与を促す．

評価指標のサポート 評価指標としては，完全一致などのルールベース手法，BLEU [21] や ROUGE [22] などの表層的な評価尺度に加えて，LLM に基づく評価尺度をサポートする．LLM に基づく評価で使用した全プロンプトは公開済みであり，透明性を担保している．

4 大規模視覚言語モデルの評価例

本節では公開されている日本語 VLM および多言語対応 VLM の日本語性能を，llm-jp-eval-mm を用いて評価を行う．

設定 VLM の推論に関して，14 種類の公開モデルは全て Hugging Face Hub[3] を通じて取得した．推論時のハイパーパラメータは表 2 のように定め，プロンプトはそれぞれのモデルが推奨するテンプレートを利用した．Heron Bench，JA-VLM-Bench-In-the-Wild では 10 回の評価スコアを平均し，残りのベンチマークでは 1 回の実行に対する評価スコアを採用した．本稿において提示する評価尺度は LLM-as-a-Judge および多肢選択式問題の正答率を利用し，それぞれを LLM，Acc. と表記している[4]．LLM-as-a-Judge において，評価者となる LLM には Azure OpenAI API 上にデプロイした gpt-4o-2024-05-13 を採用した．評価者となる LLM の temperature，seed はともに 0 とした．なお同一の seed を使用した場合でも，出力は決定的ではない[5] ことに注意が必要である．付録 A により詳細な設定を示す．

結果 llm-jp-eval-mm による評価結果を表 3 に示す．表の一段目は，日本語に特化した VLM の性能を比較している．比較対象としたモデルの中では，llm-jp-3 VILA がほとんどの評価データセットで最も高い性能を示した．特に，単一画像を入力とする質問応答課題（Heron，VGVQA，JVB-ItW）において，次点以下のモデルと 10%以上の性能差が見られた．

表の二段目では 10B パラメータ級の言語モデルを有する多言語 VLM の性能を比較している．多言語 VLM の中では，Qwen2-VL が最も優れた結果を示した．llm-jp-3 VILA や他の多言語 VLM と比較すると，複数画像を扱う MulIm-VQA，JDocQA，JMMMU などのタスクで顕著な性能差が見られる．この差異は，解像度に応じた画像処理手法（Naive Dynamic Resolution），大規模な事前学習の実施など，Qwen2-VL に導入された複数の手法による効果であると考えられる．

3） https://huggingface.co/models
4） リーダーボードにはその他のモデルの結果や，開発者の提示する尺度で計算したスコアも提示している．
5） https://platform.openai.com/docs/advanced-usage/reproducible-outputs

表 3 競争力のある視覚言語モデルの llm-jp-eval-mm を用いた日本語タスクでの評価例．“–” は評価データセットを学習に用いているためスコアが算出できないことを示す．Heron, JVB-ItW は標準偏差を併記している．

Models	Heron	JVB-ItW	VGVQA	MulIm-VQA	JDocQA		JMMMU
	LLM (%)	LLM (/5.0)	LLM (/5.0)	LLM (/5.0)	Acc.	LLM (/5.0)	Overall Acc.
Japanese InstructBLIP Alpha [10]	22.7±0.7	1.31±0.03	–	2.50	0.123	1.90	0.271
Japanese Stable VLM [13]	25.5±1.4	2.56±0.04	–	2.27	0.128	1.77	0.253
Llama-3-EvoVLM-JP-v2 [6]	39.7 ±0.3	3.23±0.06	3.17	2.90	0.152	2.23	0.357
LLaVA-CALM2-SigLIP [14]	42.6± 0.3	3.35±0.03	3.29	2.43	0.082	1.85	0.271
LLM-jp-3 VILA 14B [15]	59.9±0.4	3.77±0.02	3.68	3.38	0.175	2.45	0.285
LLaVA-1.5 7B [12]	36.3±0.4	2.56±0.02	2.74	2.07	0.145	1.98	0.296
LLaVA-1.6 7B [23]	26.4±0.2	2.44±0.03	2.72	1.89	0.140	1.75	0.255
Pangea-7B [16]	45.0±0.5	3.33±0.03	–	2.89	0.158	2.21	0.394
Pixtral-12B [24]	53.0±0.2	3.62±0.07	3.25	3.76	0.144	2.36	0.331
Llama 3.2 11B Vision Instruct [25]	33.6±1.3	2.81±0.05	3.03	2.30	0.174	2.22	0.397
InternVL2 8B [26]	46.6±0.1	3.11±0.06	3.20	2.54	0.197	2.58	0.390
Qwen2-VL 7B Instruct [18]	55.5±0.4	3.61±0.03	3.60	4.16	0.270	3.24	0.480
InternVL2 26B [26]	53.8±0.1	3.53±0.06	3.40	3.18	0.146	2.42	0.393
Qwen2-VL-72B-Instruct [18]	75.8±0.6	3.99±0.06	3.75	4.45	**0.283**	**3.66**	**0.595**
GPT-4o (detail: auto) [27]	**89.1±1.1**	**4.05±0.05**	**3.82**	**4.63**	0.239	3.60	0.566

LLM-as-a-Judge 評価スコアの信頼性 ベンチマークタスクの中でも，Heron Bench と JVB-ItW は，表 1 に示すように事例数が少ない．その結果，LLM が出力する評価スコアの非決定性に影響を受け，全体のスコアが上下しやすいと考えられる．ここでは LLM-as-a-Judge 評価の安定性を検証するため，Heron Bench および JVB-ItW について各モデル 10 回ずつ評価を行い，得られるスコアの標準偏差を測定した．その結果，表 3 が示すように，**実行ごとに Heron では約 1 ポイント，JVB-ItW では約 0.05 ポイントの変動が生じる**ことが分かった．これらの値より小さなスコア差は，統計的に有意とは言えず，LLM を用いた評価の安定性を考慮する必要性が認められる．

パラメータ数の性能向上への寄与 多言語 VLM において，言語モデル部分のパラメータ数が性能に及ぼす影響を検証するため，InternVL，Qwen2-VL を対象に，異なるパラメータ規模での性能比較を行った．表中の第二・第三段に示した評価スコアを比較すると，**言語モデルのパラメータ数が増加するにつれて性能が向上する傾向**が確認された．一方，日本語 VLM に関しては，現時点で約 10B パラメータ級の言語モデルと視覚エンコーダが統合されたモデルが主流であり，それ以上の大規模モデルを用いた VLM は公開されていない．同等のパラメータ規模で比較したとき，日本語 VLM は良好な性能を示しているが，将来的に 100B パラメータ級の言語モデルを活用した VLM 開発が進めば，さらなる性能向上が期待される．

5 おわりに

本研究では，日本語 VLM の体系的な性能評価を可能にする評価基盤ツールキット llm-jp-eval-mm を提案した．本ツールキットは公開データセットに基づく VLM の日本語性能評価の先駆的枠組みである．本ツールキットを用いた評価例から，日本の VLM 開発の進展に伴う性能向上および，LLM-jp-3 VILA, Qwen2-VL の他 VLM に対する同一パラメータ帯での日本語画像質問応答課題における優位性が観察された．その一方で GPT-4o などの大規模商用モデルが持つ視覚言語理解能力とは未だ差があることが浮き彫りとなった．

本ツールキットはマルチモーダル評価基盤の出発点であり，評価手法およびデータセット群の継続的な更新を予定している．現時点では画像質問応答タスクの充実に重点を置いているが，視覚情報への正確な接地，OCR や医療・文書といった特定領域への特化，さらには画像生成などの非テキスト的応答を評価するための多様なデータセットが必要である．さらに，今後は 3D 視覚，音声，動画，VLA（Vision-Language-Action）などのモダリティを拡張した評価基盤が考えられる．英語圏では OFA [28] や NExT-GPT [29] をはじめ非テキスト出力が可能なモデル開発が進められており，本研究の取り組みは将来的な日本語マルチモーダルモデルの多面的評価の礎となる．これら新規評価データセットの設計やモダリティの拡張を進めることも今後の課題としたい．

謝辞 本研究の成果は，JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業（博士後期課程学生支援）JPMJBS2417 及び JPMJBS2430 の支援，文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援，データ活用社会創成プラットフォーム mdx を利用して得られたものです．

参考文献

[1] Namgi Han, et al. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会 (NLP2024), 2024.
[2] Peng Xu, et al. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. arXiv:2306.09265, 2023.
[3] Kaichen Zhang, et al. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models. arXiv:2407.12772, 2024.
[4] Yuichi Inoue, et al. Heron-Bench: A Benchmark for Evaluating Vision Language Models in Japanese. arXiv:2404.07824, 2024.
[5] Inoue Yuichi, et al. JA-Multi-Image-VQA. https://huggingface.co/datasets/SakanaAI/JA-Multi-Image-VQA, 2024.
[6] Takuya Akiba, et al. Evolutionary Optimization of Model Merging Recipes. arXiv:2403.13187, 2024.
[7] Nobuyuki Shimizu, et al. Visual Question Answering Dataset for Bilingual Image Understanding: A Study of Cross-Lingual Transfer Using Attention Maps. In **Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)**, pp. 1918–1928, 2018.
[8] Eri Onami, et al. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9503–9514, 2024.
[9] Shota Onohara, et al. JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark for Culture-aware Evaluation. arXiv:2410.17250, 2024.
[10] Makoto Shing and Takuya Akiba. Japanese Instruct-BLIP Alpha. https://huggingface.co/stabilityai/japanese-instructblip-alpha, 2023.
[11] Haotian Liu, et al. Visual Instruction Tuning. In **Advances in Neural Information Processing Systems 36 (NeurIPS 2023)**, pp. 34892–34916, 2023.
[12] Haotian Liu, et al. Improved Baselines with Visual Instruction Tuning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)**, pp. 26296–26306, June 2024.
[13] Makoto Shing and Takuya Akiba. Japanese Stable VLM. https://huggingface.co/stabilityai/japanese-stable-vlm, 2024.
[14] Aozora Inagaki. LLaVA-CALM2-SigLIP. https://huggingface.co/cyberagent/llava-calm2-siglip, 2024.
[15] Keito Sasagawa, et al. Constructing Multimodal Datasets from Scratch for Rapid Development of a Japanese Visual Language Model. arXiv:2410.22736, 2024.
[16] Xiang Yue, et al. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages. arXiv:2410.16153, 2024.
[17] Hanoona Rasheed, et al. Palo: A Large Multilingual Multimodal Language Model. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)**, 2025.
[18] Peng Wang, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191, 2024.
[19] Percy Liang, et al. Holistic Evaluation of Language Models. **Transactions on Machine Learning Research**, 2023.
[20] Leo Gao, et al. A Framework for Few-shot Language Model Evaluation. =https://zenodo.org/records/12608602, 2024.
[21] Kishore Papineni, et al. BLEU: A Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)**, pp. 311–318, 2002.
[22] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In **Text Summarization Branches Out**, pp. 74–81, 2004.
[23] Haotian Liu, et al. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024.
[24] Pravesh Agrawal, et al. Pixtral 12B. arXiv:2410.07073, 2024.
[25] Meta. Llama-3.2-11B-Vision. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision, 2024.
[26] Zhe Chen, et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. arXiv:2404.16821, 2024.
[27] OpenAI. GPT-4 Technical Report. arXiv:2303.08774, 2023.
[28] Peng Wang, et al. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. arXiv:2202.03052, 2022.
[29] Shengqiong Wu, et al. NExT-GPT: Any-to-Any Multimodal LLM. In **Proceedings of the 41st International Conference on Machine Learning (ICML 2024)**, pp. 53366–53397, 2024.
[30] Lianmin Zheng, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)**, 2024.

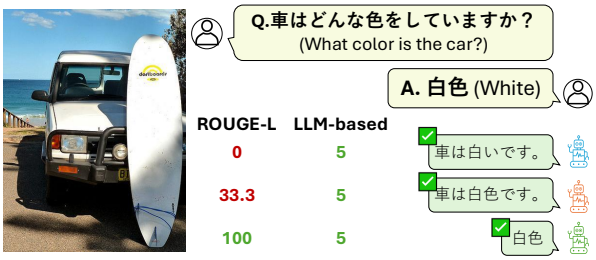

図 2　日本語 VQA の回答を評価する際，ROUGE-L スコアが人手評価と合致しない例．

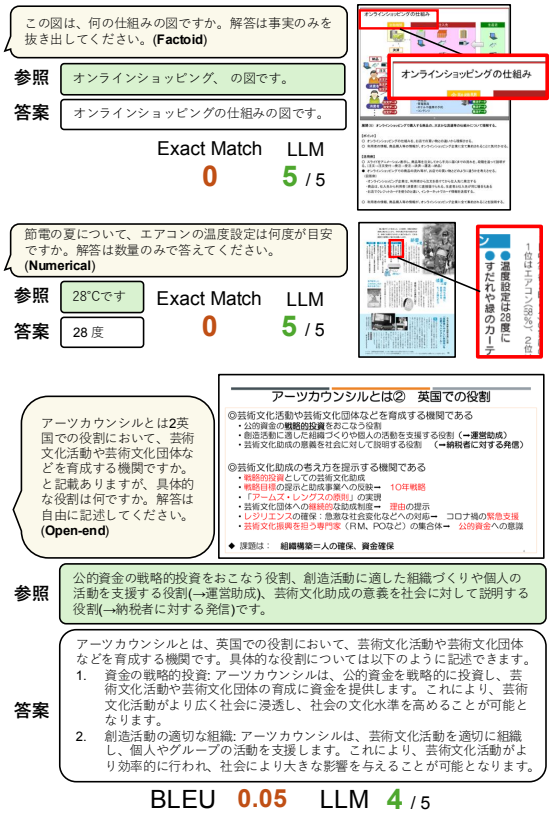

図 3　JDocQA におけるデフォルト評価指標の不都合．

付録

A　評価設定の詳細

Heron Bench の評価スコアは，GPT-4 を基準とした相対評価によって与えられる．評価対象となるモデルの評価スコアは，全サンプルの得点の平均を GPT-4 の平均スコアで割った百分率で算出される．そのため，値が 100 を超えうることに留意せよ．

JA-VLM-Bench-In-the-Wild，JA-VG-VQA500，および JA-MultiImage-VQA では，評価指標として ROUGE-L [22] が一般的に使用される．しかし，こと日本語質問応答においては，回答の文体によって ROUGE-L の値は大きく変化する．正しい内容を含んでいても，異なる文体で生成された文は低いスコアになる可能性がある．

図 2 は文章構造や表現の違いがスコアに影響する典型的な例である．3 つの異なる回答はすべて正しいが，“白色” 以外の回答の ROUGE-L スコアは 0 である．このような例を鑑み，日本語の微妙な意味や文体変化を捉える上で，現状で ROUGE-L は VLM の日本語応答性能を正確に評価できていないと判断した．過小評価は多数の参照回答を用意することで抑制できるが，可能な回答を網羅することは困難であるため，我々は LLM-as-a-Judge での評価を行った．GPT-4o を評価者として，標準的な画像質問応答のプロンプトテンプレート[6] にいくつかの修正を加えた上で与え，生成された回答を 5 段階のリッカート尺度で評価させた．

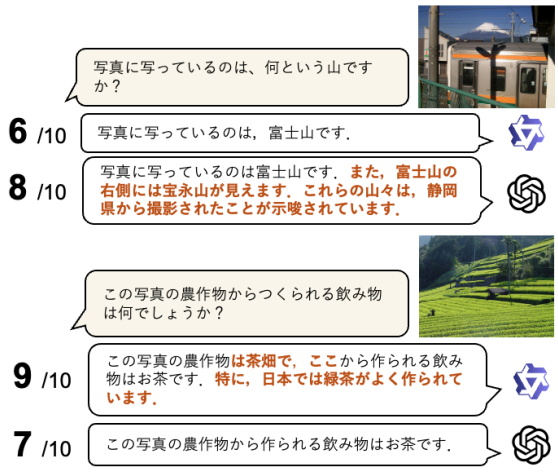

図 4　Qwen2-VL 72B と gpt-4-0125-preview の出力から見られる LLM-as-a-Judge の評価バイアス．

JDocQA は LLM を訓練するための学習セットがあるが，評価実験では学習セットを利用して微調整したモデルの評価は行わなかった．評価指標に関して，著者は Factoid，Numerical カテゴリの回答を完全一致で評価することを推奨しているが，図 3 に示すように，文体によって間違いであると誤認される例が散見された．そのため LLM-as-a-Judge を用いて正誤判定を行った．

また，Open-end カテゴリの質問応答を BLEU [21] で評価している．しかし GPT-4o をはじめとする直近のモデルは人間の選好に適応した特定のスタイルで回答する．前述した ROUGE-L と同様，簡潔な回答をするモデルと比べて不利な評価設定となる．そこで我々は，LLM を用いた 5 段階のリッカート尺度での評価を行い，代表的な評価指標として表 3 に提示した．今後 JDocQA 上での性能を BLEU や完全一致で評価する場合，参照回答の再アノテーションが必要である．

B　LLM-as-a-Judge の評価バイアス

llm-jp-eval-mm の評価に LLM-as-a-Judge アプローチを採用しているが，評価者となる LLM は人間と異なる選好を示す．その例のひとつに，**冗長な回答を不当に高く評価する傾向**がある．図 4 のように，実際の回答として与えるべき情報に加えて，明らかに冗長な情報を含んだ回答を高く評価している．この傾向は先行研究 [30] でも報告されており，日本語 VLM の出力に対する評価においても同様であることが確認された．これ以外にも未知のバイアスを含むおそれがあることに留意せよ．

6) https://cloud.google.com/vertex-ai/generative-ai/docs/models/metrics-templates#pointwise_question_answering_quality